Visualisation and Labelling of Clusters

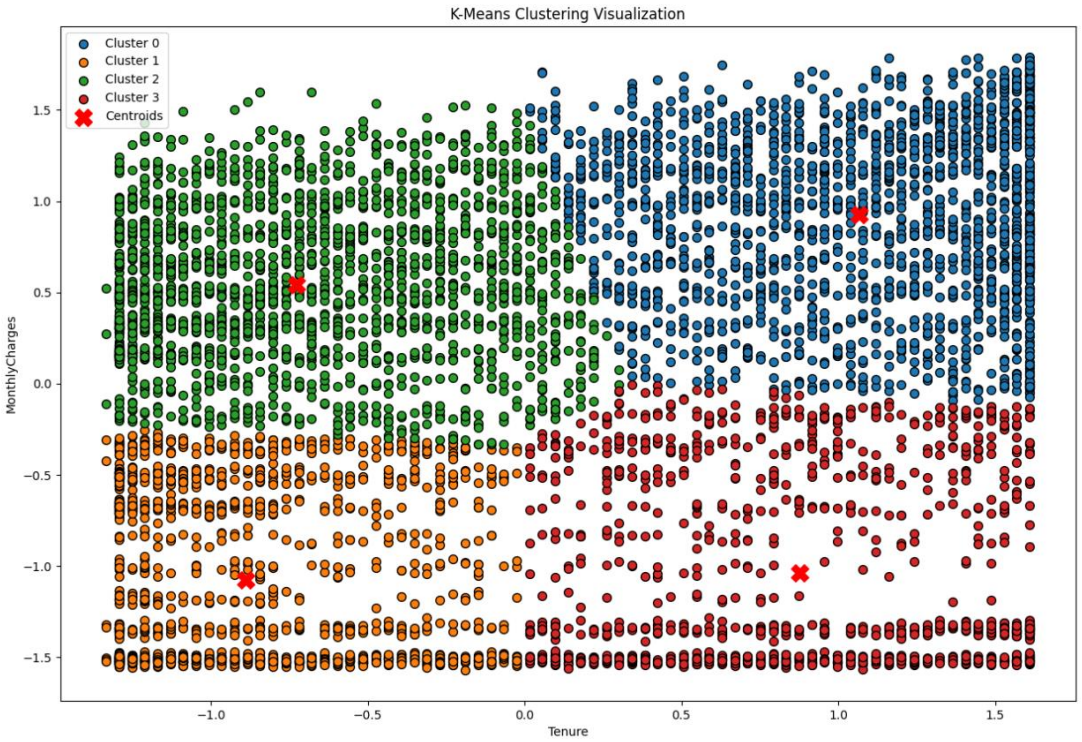
Code:

We are using **matplotlib.pyplot** - a submodule of the Matplotlib library in Python, which provides a collection of functions that make plotting and data visualization easy. In particular, we use matplotlib.pyplot.scatter() to create a scatter plot for our two features: **tenures** and **MonthlyCharges**.

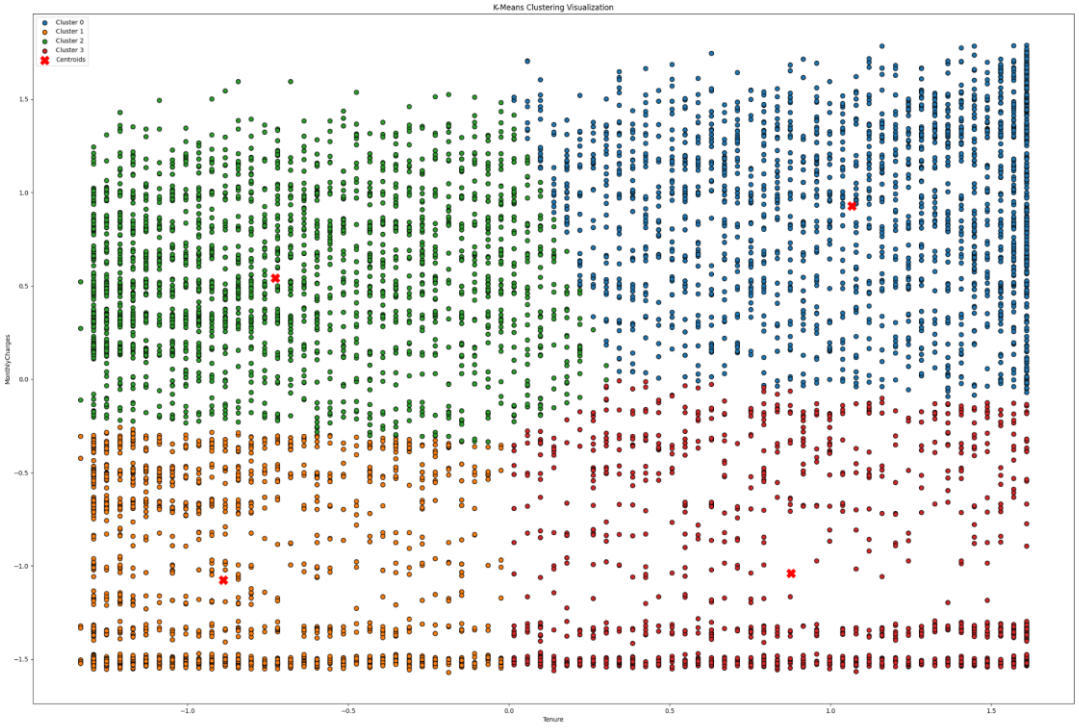
- A scatter plot is a type of plot that displays individual data points as dots (markers) on a two-dimensional graph, where each point represents the values of two variables (one plotted along the x-axis and the other on the y-axis).

```
# Visualize model
import matplotlib.pyplot as plt
import numpy as np

# Predict cluster labels for the data
cluster_labels = kmeans.predict(scaled_features)
# Access the cluster centroids
centroids = kmeans.cluster_centers_

# Scatter plot of data points with cluster labels
plt.figure(figsize=(15, 10 ))

# Unique cluster labels
unique_labels = np.unique(cluster_labels)
# Loop through each unique cluster label and plot the data points for that cluster
for cluster in unique_labels:
    # Select data points that belong to the current cluster
    cluster_data = scaled_features[cluster_labels == cluster]

    # Scatter plot for the current cluster
    plt.scatter(cluster_data[:, 0], cluster_data[:, 1],
                label=f'Cluster {cluster}', s=50, edgecolor='k', marker='o')

plt.scatter(centroids[:, 0], centroids[:, 1], s=200, c='red', marker='X', label='Centroids')
plt.title('K-Means Clustering Visualization')
plt.xlabel('Tenure')
plt.ylabel('MonthlyCharges')
plt.legend()
plt.show()
```

Output:



Original scatter plot with ***figsize = (15,10)***



Same data, same scatter plot, but with larger figure size - ***figsize = (30,20)***

Detailed Analysis of Scatter Plot

1. **Cluster Distribution:**
    - Cluster 0 (Blue):
        - These data points are spread across a wide range of tenures (from around 0.5 to 1.5 on the x-axis).
        - They also have high monthly charges (from about 0.0 to 1.5 on the y-axis).
        - This cluster likely represents customers with **long tenures** and **high monthly charges**.
    - Cluster 1 (Orange):
        - These points are concentrated towards the lower-left of the plot, with tenures ranging from around -1.5 to 0 and lower monthly charges (from -1.0 to -1.5 on the y-axis).
        - This suggests that Cluster 1 represents customers with **short tenures** and **low monthly charges**.
    - Cluster 2 (Green):
        - These points are spread across the middle, representing a wide range of tenures (from around -1.5 to 0.5).
        - Monthly charges range from 0 to 1.0.
        - Cluster 2 likely represents customers with **moderate tenures** and **moderate monthly charges**.
    - Cluster 3 (Red):
        - These points are mostly concentrated on the right side of the plot (tenure around 0 to 1.5) with lower monthly charges (from -1.5 to 0).
        - This cluster seems to represent customers with **longer tenures** but **lower monthly charges**.
2. **Cluster Separation**:
    - The clusters seem reasonably well-separated, especially between Clusters 0 (Blue) and Cluster 1 (Orange). There is some overlap between Clusters 2 (Green) and Cluster 3 (Red), particularly in the middle range of tenures, which suggests there might be some similarity between customers in these groups.
    - Cluster 2 (Green) and Cluster 3 (Red) overlap a bit around the center (around tenures between -0.5 to +0.5). This indicates some customers have similar tenures but different monthly charges, which might explain the overlap.
3. **Centroid Positions**:
    - The centroids for each cluster are fairly distinct, located roughly at the mean positions of the data points.

- The Blue cluster's centroid is further to the right (representing longer tenures and higher monthly charges).
- The Orange cluster's centroid is closer to the bottom-left, representing customers with shorter tenures and lower monthly charges.
- Green and Red centroids are more centralized, reflecting moderate tenures and a wider range of monthly charges.

Interpretation of Clusters Based on Scatter Plot

1. **Cluster 0 (Blue)**:
   - Customers in this cluster have **longer tenures** and are paying **higher monthly charges**. This group could represent loyal, high-paying customers, and we might want to focus on **customer retention strategies** here.
2. **Cluster 1 (Orange)**:
   - These are **short-tenure, low-charge customers**. This group may represent **new customers** or customers on **lower-priced plans**. We could potentially target this group with **upsell or retention strategies**.
3. **Cluster 2 (Green)**:
   - Customers in this cluster have **moderate tenures** and are paying **moderate monthly charges**. This group likely represents **average customers** who may not require special retention strategies but could be further segmented or analysed.
4. **Cluster 3 (Red)**:
   - Customers in this cluster have **longer tenures** but **lower monthly charges**. These might be **loyal customers** who are sticking around but are on **lower-priced plans**. Consider **cross-selling or offering higher-tier services** to this group to increase their revenue potential.

Actionable Insights

Targeted marketing or retention campaigns:

- Cluster 0: Focus on **retention and loyalty programs** to keep these high-value customers happy.
- Cluster 1: Potential for **upselling or cross-selling**, as they are newer customers with lower spending.
- Cluster 3: Since they have long tenures but pay lower monthly charges, there might be opportunities to **introduce higher-priced plans** or premium features.

Deeper Analysis of Each Cluster

We can perform a deeper dive into the characteristics of each cluster by analyzing the original (unscaled) values of the features ("tenure" and "MonthlyCharges") rather than the standardized values used during clustering. This allows us to better understand what each cluster represents in terms of the real-world data, which is more interpretable.

```
df = pd.DataFrame(data_encoded_cleaned[['tenure', 'MonthlyCharges']], columns=['tenure', 'MonthlyCharges'])
```

Original DataFrame contains two features (before scaling)

```
# Predict cluster labels for the data
cluster_labels = kmeans.predict(scaled_features)

# Add a Cluster column to the original DataFrame, this column contains label for each data point
df['Cluster'] = cluster_labels
# Group by the 'Cluster' column
cluster_summary = df.groupby('Cluster').agg({
    'tenure': ['mean', 'median', 'min', 'max'],
    'MonthlyCharges': ['mean', 'median', 'min', 'max']
})
# Display the summary statistics
print(cluster_summary)
```

Output:

	tenure				MonthlyCharges			
	mean	median	min	max	mean	median	min	max
Cluster								
0	58.688272	61.0	33	72	92.972222	94.65	62.50	118.75
1	10.879331	8.0	0	32	33.072401	25.40	18.25	57.55
2	14.822558	13.0	0	40	81.395828	80.25	55.25	112.95
3	54.075372	54.0	33	72	34.127432	25.20	18.40	65.00

Interpretation of Clusters Based on original value

Cluster 0:

- Contains customers who have **long tenures** (33 to 72 months), with an average of **58.69 months**.
- These customers pay **high monthly charges**, with a mean of **$92.97**.
- This cluster likely represents **long-term, high-paying customers**. These customers have been with the company for a while and are on **premium or higher-tier plans**.

Cluster 1:

- Consists of customers with **short tenures** (0 to 32 months), with an average of **10.88 months**.
- These customers are paying **low monthly charges**, with a mean of **$33.07**.
- This cluster likely represents **newer customers on lower-priced plans** or customers with **limited engagement**.

Cluster 2:

- Has customers with **moderate tenures** (0 to 40 months), with an average of **14.82 months**.
- These customers are paying **high monthly charges** (mean of **$81.40**).
- This group likely consists of **newer customers who have opted for premium plans** or mid-tenure customers with **higher spending habits**.

Cluster 3:

- Includes customers with **longer tenures** (33 to 72 months), with an average of **54.08 months**.
- Despite their long tenure, these customers are paying **low monthly charges** (mean of **$34.13**).
- This group likely represents **loyal customers on lower-tier plans**.

Overall Insights:

- Cluster 0 and Cluster 3 represent long-term customers, but Cluster 0 is paying significantly more than Cluster 3.
- Cluster 1 and Cluster 2 represent shorter-term customers, but Cluster 2 is paying higher monthly charges, which might indicate a preference for premium services despite shorter engagement.

Suggested Actions:

1. Cluster 0 (High Tenure, High Charges):
    - Focus on retention and loyalty programs. These are high-value customers, so maintaining their satisfaction is crucial.
2. Cluster 1 (Low Tenure, Low Charges):
    - Upsell and cross-sell strategies to move these customers to higher-paying plans.
    - Focus on early engagement to prevent churn.
3. Cluster 2 (Moderate Tenure, High Charges):
    - Focus on retention efforts, as these are potentially high-value customers who might be at risk of churn due to shorter tenures.
4. Cluster 3 (High Tenure, Low Charges):
    - Consider upselling and offering incentives to move these loyal customers to higher-paying plans.